DATA MINING TO DATABASES:A REVIEW PAPER

[1]Ayush Kumar, [2]Neelesh Jain, [3]Neeraj Gupta
[1]Student, [2]Professor, [3]Professor
[1,2,3] Department of Computer Science & Engineering
[1,2,3] SAM College of Engineering and Technology, Bhopal

Abstract:- The modern era has witnessed a rapid surge in data generation, prompting the need for advanced analytical methods to efficiently process, analyze, and extract value from large volumes of data. In this context, this study aims to integrate data mining techniques into database systems to bolster their analytical capabilities and streamline decision-making processes. We focus on five key aspects: preprocessing, clustering, classification, association rule mining, and anomaly detection. Preprocessing ensures data quality and consistency, eliminating noise, and handling missing values. Clustering groups similar data points based on their attributes, facilitating pattern recognition and data segmentation. Classification categorizes data into predefined classes, enabling predictive modeling and improved understanding of relationships among data points. Association rule mining identifies frequent itemsets and generates rules to uncover relationships among variables, supporting business intelligence and decision-making.

Keywords: Preprocessing, Clustering, Classification, Association Rule Mining, Anomaly Detection

I. INTRODUCTION

Data mining and knowledge discovery indatabases have been attracting a significant amount of research, industry, and media attention of late. What is all the excitement about? This article provides an overview of this emerging field, clarifying how data mining and knowledge discovery in databases are related both to each other and to related fields, such as machine learning, statistics, and databases. The article mentions particular real-world applications, specific data-mining techniques, challenges in- volved in real-world applications of knowledge discovery, and current and future search directionsin the field.

An abstract level, the KDD field is concerned with the development of methods and techniques for making sense of data. The basic problem addressed by the KDD process is one of mapping low-level data (which are typically too volume in digest easily) into other forms that might be more compact (for example, a short report), more ab-stract (for example, a descriptive approximation or model of the process that generated the data),or more useful(for example, a predictive model for estimating the value of future cases). At the core of the processing the application of specific data- mining methods for pattern discovery and extraction.[1]

This article begins by discussing the historical context of KDD and data mining and their intersection with other related fields. A brief summary of recent KDD real-world applications is provided. Definitions of KDD and data mining are provided, and the general multistep KDD process is outlined. This multistep process has the application of data-mining algorithms as one particular step in the process. The data-miningstep is discussed in more detail in the context of specific data-mining algorithms and their application. Real-world practical application issues are also outlined. Finally, the article enumerates challenges for future research and development and in particular discusses potential opportunities for AI technology in KDD systems.

Intimately familiar with the data and serving as an interface between the data and the users and products.

For these (and many other) applications, this form of manual probing of a data set is slow, expensive, and highly subjective. Infact, as data volumes grow dramatically, this type of manual data analysis is becoming completely impractical in many domains. Databases are increasing in size intwo ways:

(1) the number N of records or objects in the database and (2) the number d of fields or at-tributes to an object. Databases containing on the order of $N = 10^9$objects are becoming in- creasingly common, for example ,in the as-tronomical sciences. Similarly, the number of field sd can easily be on the order of 10^2or even 10^3, for example, in medical diagnostic applications. Who could be expected to digest millions of records, each having tensor hundreds of fields? We believe that this job is certainly not one for humans; hence, analysis work needs to be automated, at least partially. The need to scale up human analysis capabilities to handling the large number of bytes that we can collect is both economic and scientific. Businesses use data to gain competitive advantage, increase efficiency, and provide more valuable services to customers. Data we capture about our environment are the basic evidence we use to build theories and models of the universe we live in. Because computers have enabled humans to gather more data than we can digest, it is onlynatural to turn to computational

techniques to help us unearth meaningful patterns and structures from them assive volumes of data. Hence, KDD is an attemptto address a problem that thedigitalin formation eramadea fact of life for all of us:data overload.

DATA MINING AND KNOWLEDGE DISCOVERY IN THE REAL WORLD

A large degree of the current interest in KDD isthe result of the media interest surrounding successful KDD applications, for example, the focus articles within the last two years in *Business Week*, *Newsweek*, *Byte*, *PC Week*, and other large- circulation periodicals. Unfortunately, it is not always easy to separate fact from media hype. Nonetheless, several well-documented examples of successful systems can rightly be referred to as KDD applications and have been deployed in operational use on large-scalereal-world problem s in science and in business.

In science, one of the primary application areas is astronomy. Here, a notable success was achieved by SKICAT, a system used by astronomers to perform image analysis, classification, and cataloging of sky objects from sky-survey images (Fayyad, Djorgovski, and Weir 1996). In its first application, the system was used to process the 3 terabytes (10^{12}bytes) of image data resulting from the Second Palomar Observatory Sky Survey, where it is estimated that on the order of 10^9sky objects are detectable. Skicat can outer- form humans and traditional computational techniques in classifying faint sky objects. See Fayyad, Haussler, and Stolorz (1996) for a survey of scientific applications.

In business, main KDD application areas includes marketing, finance (especially investment), fraud detection, manufacturing, telecommunications, and Internet agents.

Marketing: In marketing, the primary application is database marketing systems, which analyze customer databases to identify different customer groups and forecast their behavior. *Business Week* (Berry 1994) estimated that over half of all retailers are using or planning to use database marketing, and those who do use it have good results; for example, American Express reports a 10- to 15-percent increase in credit-card use. Another notable marketing application is market- bas-ket analysis (Agrawal et al. 1996) systems, which find patterns such as, "If customer bought X, he/she is also likely to buy Y andZ." Such patterns are valuable to retailers.

Investment: Numerous companies use data mining for investment, but most do not describe their systems. One exception is LBS Capital Management. Its system uses expert systems, neuralnets, and genetic algorithms to manage portfolios totaling $600 million; since its start in 1993, the system has outperformed the broad stock market (Hall, Mani, and Barr 1996).

Fraud detection: HNC Falcon and Nestor PRISM systems are used for monitoring credit- card fraud, watching over millions of accounts. The FAIS system (Senator et al. 1995), from the U.S. Treasury Financial Crimes Enforcement Network, is used to identify financial transactions that might indicate money-laundering activity.

Manufacturing: The Cassiopee troubleshooting system, developed as part of a joint venture between General Electric and Snecma, was applied by three major European airlines to diagnose and predict problems for the Boeing 737. To derive families of faults, clustering methods are used. Cassiopee received the European first prize for innova- applications (Manago and Auriol 1996).

Telecommunications: The telecommunications alarm-sequence analyzer (TASA) was built in cooperation with a manufacturer of telecommunications equipment and three telephone networks (Mannila, Toivonen, and Verkamo 1995). The system uses a novel framework for locating frequently occurring alarm episodes from the alarm stream and presentingthem as rules. Large sets of discovered rules can be explored with flexible information-retrieval tools supporting interactivity and iteration. In this way, TASA offers pruning, grouping, and ordering tools to refine the results of a basic brute-force search for rules.

Data cleaning: The Merge-Purge system was applied to the identification of duplicate welfare claims (Hernandez and Stolfo 1995). It was used successfully on data from the Welfare Department of the State of Washington.

In other areas, a well-publicized system is IBM's Advanced Scout, a specialized data-mining system that helps National Basketball Association (NBA) coaches organize and interpret data from NBA games (U.S. News 1995). Advanced Scout was used by several of the NBA teams in 1996, including the Seattle Su-personas, which reached the NBA finals.

Finally, a novel and increasingly important type of discovery is one based on the use of intelligent agents to navigate through an information-rich environment. Although the idea of active triggers has long been analyzed in the database field, really successful applications of this ideaappeared only with the advent of the Internet. These systems ask the user to specify profile of interest and search for related information among a wide variety of public-do-main and proprietary sources. For example, FIREFLY is a personal music-recommendation agent: It asks a userhis/her opinion of several music pieces and then suggests other music that the user might like (<http:// www. ffly. com/>).CRAYON (http://crayon.net/>) allows users to create the irown free news paper (supported by ads); Newshound (<http://www. sjmercury. com/ hound/>) from the *San Jose Mercury News* and Far cast (<http://www. far- cast. com/> automatically search information from a wide variety of sources, including newspapers and wire services, and e-mail relevant documents directlyto the user.

These are just a few of the numerous such Systems that use KDD techniques to automat- ically produce useful information from large masses of raw data. See Piatetsky-Shapiro et al. (1996) for an overview of issues in developing industrial KDD applications.

DATA MINING AND KDD

Historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. The term *data mining* has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities. It has also gained popularity in the database field. The phrase *knowledge discovery in databases* was coined at the first KDD workshop in 1989 (Piatetsky-Shapiro 1991) to emphasize thatknowledge is the end product of a data-drivendiscovery. It has been popularized in the AI and machine-learning fields.

In our view, KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. *Data mining* is the application of specific algorithms for extracting patterns from data. The distinction between the KDD process and the data-mining step (within the process) is a central point of this article. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data. Blind application of data-mining methods (rightly criticized as data dredging in the statistical literature) can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns.

Related AI research fields include machine discovery, which targets the discovery of em- pirical laws from observation and experimentation (Shrager and Langley 1990) (see Kloes gen and Zytkow [1996] for a glossary of terms common to KDD and machine discovery),and causal modeling for the inference of causal models from data (Spirtes, Glymour, and Scheines 1993). Statistics in particular has much in common with KDD (see Elderand Pregibon [1996] and Glymour et al. [1996] for a more detailed discussion of this synergy). Knowledge discovery from data is fundamentally a statistical endeavor. Statistics provides a language and frame work for quan- tifying the uncertainty that results when one tries to infer general patterns from a particular sample of an overall population. As mentioned earlier, the term *data mining* has had negative connotations in statistics since the 1960s when computer-based data analysis techniques were first introduced. The concern arose because if one searches long enough in any data set (even randomly generated data), one can find patterns that appear to be statistically significant but, in fact, are not. Clearly, this issue is of fundamental importance to KDD. Substantial progress has been made in recent years in understanding such issues in statistics. Much of this work is of direct relevance to KDD. Thus, data mining is a legitimate activity as long as one understands how to do it correctly; data mining carried out poorly (without regard to the statistical aspects of the problem) is to be avoided. KDD can also be view edasen compassing a broader view of modeling than statistics. KDD aims to provide tools to automate (to the degree possible) the entire process of data analysis and the statistician’s “art” of hypothesis selection.

A driving force behind KDD is the database field (the second D in KDD). Indeed, the problem of effective data manipulation when data cannot fit in the main memory is of fundamental importance to KDD. Database techniques for gaining efficient data access, grouping and ordering operations when accessing data, and optimizing queries constitute the basics for scaling algorithms to larger data sets. Most data-mining algorithms from statistics, pattern recognition, and machine learning assume data are in the main memory and pay no attention to how the algorithm breaks down if only limited views of the data are possible.

A related field evolving from databases is *data ware housing*, which refers to the popular business trend of collecting and cleaning transactional data to make them available for online analysis and decision support. Data warehousing helps set the stage for KDD in two important ways: (1)data cleaning and (2) data access.

Data cleaning: As organizations are forced to think about a unified logical view of the wide variety of data and databases they possess, they have to address the issues of mapping data to a single naming convention, uniformly representing and handling missing data, and handling noise anderrors when possible.

Data access: Uniform and well-defined methods must be created for accessing the data and providing access paths to data that were historically difficult to get to (for example, stored offline).

Once organizations and individuals have solved the problem of how to store and access their data, the natural next step is the question, what else do we do with all the data? This is where opportunities for KDD naturally arise.

A popular approach for analysis of data warehouses is called online analytical processing (OLAP), named for a set of principles proposed by Codd (1993). OLAP tools focus on providing multidimensional data analysis, which is superior to SQL in computing summaries and breakdowns along many dimensions. OLAP tools are targeted toward simplifying and supporting interactive data analysis, but the goal of KDD tools is to automate as much of the process as possible. Thus, KDD is a step beyond what is currently supported by most standard database systems.

Basic Definitions

KDD is the non-trivial process of identifying valid, novel, potentially useful, and ultimate-

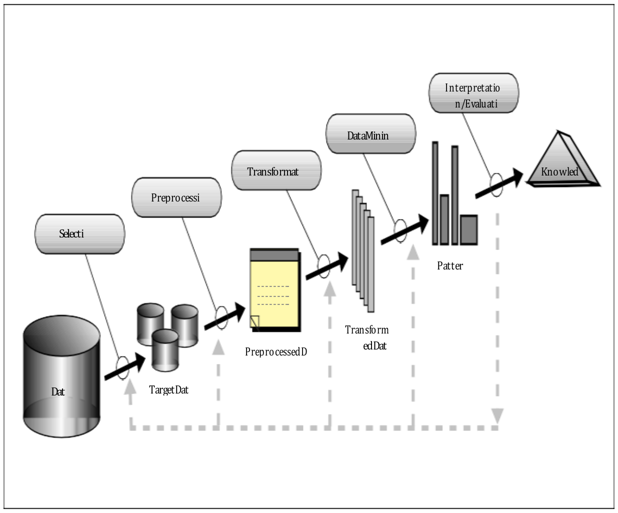


Figure 1: An Overview of the Steps, That Compose the KDD Process

Here, *data* are a set of facts (for example, cases in a database), and *pattern* is an expression in some language describing a subset of the data or a model applicable to the subset. Hence, in our usage here, extracting a pattern also designates fitting a model to data; finding structure from data; or, in general, making any high-level description of a set of data. The term *process* implies that KDD comprises many steps, which involve data preparation, search for patterns, knowledge evaluation, and refinement, all repeated in multiple iterations. By *nontrivial*, we mean that some search or inference is involved; that is, it is not a straightforward computation of predefined quantities like computing the average value of a set of numbers.

The discovered patterns should be valid on new data with some degree of certainty. We also want patterns to be novel (at least to the system and preferably to the user) and potentially useful, that is, lead to some benefit to the user or task. Finally, the patterns should be understandable, if not immediately then after some post processing.

The previous discussion implies that we can define quantitative measures for evaluating extracted patterns. In many cases, it is possible to define measures of certainty(for example, estimated prediction accuracy on new data) or utility (for example, gain, perhaps in dollars saved because of better predictions or speedup in response time of a system). Notions such as novelty and understandability are much more subjective. In certain contexts, understandability can be estimated by simplicity (for example, the number of bits to describe a pattern). An important notion, called *interestingness* (for example, see Silberschatz and Tuzhilin[1995] and Piatetsky- Shapiro and Matheus [1994]), is usually taken as an overall measure of pattern value, combining validity, novelty, usefulness, and simplicity. Interestingness functions can be defined explicitly or can be manifested implicitly through an or-dering placed by the KDD system on the dis- covered patterns or models. patterns is often infinite, and the enumeration of patterns involves some form of search in this space. Practical computational constraints place severe limits on the subspace that can beexplored by a data-mining algorithm.

The KDD process involves using the database along with any required selection, preprocessing, sub sampling, and transformations of it; applying data-mining methods (algorithms) to enumerate patterns from it; and evaluating the products of data mining to identify the subset of the enumerated patterns deemed knowledge. The data-mining component of the KDD process is concerned with the algorithmic means by which patterns are extracted and enumerated from data. The overall KDD process (figure 1) includes the evaluation and possible interpretation of the mined patterns to determine which patterns can be considered new knowledge. The KDD process also includes all the additional steps described in the next section.

The notion of an overall user-driven process is notunique to KDD: analogous proposals have been put forward both in statistics (Hand 1994) and in machine learning (Brodley and Smyth 1996).

The KDD Process

The KDD process is interactive and iterative, involving numerous steps with many decisions made by the user. Brachman and Anand (1996) give a practical view of the KDD process, emphasizing the interactive nature of the process. Here, we broadly outline some of its basic steps:

First is developing an understanding of the application domain and the relevant priorknowledge and identifying the goal of the KDD process from the customer's viewpoint.

Second is creating a target data set: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to beperformed.

Third is data cleaning and preprocessing. Basic operations include removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes.

Fourth is data reduction and projection: finding useful features to represent the data depending on the goal of the task. With dimensional education or transformation methods, the effective number of variables under consideration can be reduced, or invariant representations for the data can be found.

The Data-Mining Step of the KDD Process

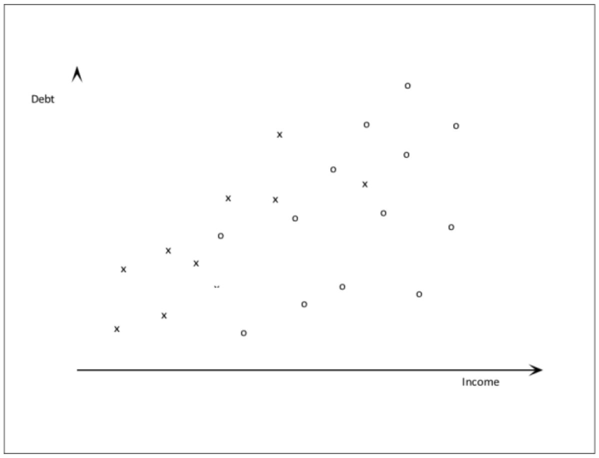
The data-mining component of the KDD process often involves repeated iterative application of particular data-mining methods. This section presents an overview of the primary goals of data mining, a description of the methods used to address these goals, and a brief description of the data-mining algorithms that incorporate thesemethods.

The knowledge discovery goals are defined by the intended use of the system. We can distinguishtwo types of goals:(1)verification and (2) discovery. With *verification,* the system is limited to verifying the user's hypothesis. With *discovery,* the system autonomously finds new patterns. We further subdivide the discovery goal into *prediction,* where the system finds patterns for predicting the future behavior of some entities, and *description,* where the system finds patterns for presentation to a user in a human-understandable form. In this article, we are primarily concerned with discovery-oriented data mining.

Data mining involves fitting models to, or determining patterns from, observed data. The fitted models play the role of inferred knowledge: Whether the models reflect useful or interesting knowledge is part of the overall, interactive KDD process where subjective human judgment is typically required. Two primary mathematical for malisms are used in model fitting: (1) statistical and (2) logical. The *statistical approach* allows for nondeterministic effects in the model, whereas a *logical model* is purely deterministic. We focus primarily on the statistical approach to data mining, which tends to be the most widely used basis for practical data-mining applications given the typical presence of uncertainty in real-world data-generating processes.

Most data-mining methods are based on tried and tested techniques from machine learning, pattern recognition, and statistics: classification, clustering, regression, and so on. The array ofdifferent algorithms under each of the seheadings can often be bewildering to both the novice and the experienced data analyst. It should be emphasized that of the many data-mining methods advertised in the literature, there are really only a few fundamental techniques. The actual underlying modeler presentation being used by a particular method typically comes from a composition of a small number of well-known op-tions: polynomials, splines, kernel and basis functions, threshold-Boolean functions, and soon. Thus, Algorithms tend to differ primarily. In the goodness-of-fit criterion used to evaluate model fit or in the search method used to find a good fit.

Figure 2: A Simple Data Set with Two Classes used



for Illustrative Purposes

In our brief overview of data-mining methods, we try in particular to convey the notion that most (if not all) methods can be viewed as extensions or hybrids of a few basic techniques and principles. We first discuss the primary methods of data mining and then show that the data- mining methods can be viewed as consisting of three primary algorithmic components: (1) model representation, (2) model evaluation, and (3) search. In the discussion of KDD and data-mining methods, we use a simple example to make some of the notions more concrete. Figure 2 shows a simple two-dimensional artificial data set consist- ing of 23 cases. Each point on the graph rep-resents a person who has been given a loan by a particular bank at some time in the past. The horizontal axis represents the income of the person; the vertical axis represents the total personal debt of the person (mortgage, car payments, and so on). The data have been classified into two classes: (1) the x's represent persons who have defaulted on their loans and (2) the o's represent persons whose loans are in good status with the bank. Thus, this simple artificial data set could represent a historical data set that can contain useful knowledge from the point of view of the bank making the loans. Note that in actual KDD applications, there are typically many more dimensions (as many as several hundreds) and many more data points (many thousands or even millions).

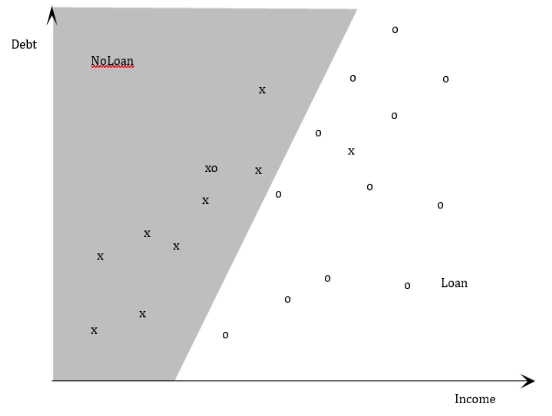


Figure 3: A Simple Linear Classification Boundary for the Loan Data Set

Data-Mining Methods

The two high-level primary goals of data mining in practice tend to be prediction and description. As stated earlier, prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest, and description focuses on finding human-interpretable patterns describing the data. Although the boundaries between prediction and description are not sharp (some of the predictive models can be descriptive, to the degree that they are understandable, and vice versa), the distinction is useful for understanding the overall discovery goal. The relative importance of prediction and description for particular data- mining applications can vary considerably. The goals of prediction and description can be achieved using a variety of particular data-mining methods.

Classification is learning a function that maps (classifies) a data item into one of severalpredefined classes (Weiss and Kulikowski 1991; Hand 1981). Examples of classification methods used as part of knowledge discovery applications include the classifying of trends in financial markets(Apteand Hong1996) and the automated identification of objects of interest in large image databases (Fayyad, Djorgovski, and Weir 1996). Figure 3 shows a simple partitioning of the loan data into two class regions; note that it is notpossible to separate the classes perfectly using a linear decision boundary. The bank might want to use the classification regions to automatically decide whether future loan applicants will be given a loan or not.

Regression is learning a function that maps a data item to a real-valued prediction variable. Regression applications are many, for example, predicting the amount of biomass present in a forest given remotely sensed microwave measurements, estimating the probability that apatient will survive given the results of a set of diagnostic tests, predicting consumer demand for a new product as a function of advertising expenditure, and predicting time series where the input variables can be time-lagged versions of the prediction variable. Figure 4 shows the result of simple linear regression where total debt is fitted as a linear function of income: The fit is poor be- cause only a weak correlation exists between the two variables.

High dimensionality: Not only is there often a large number of records in the database, but there can also be a large number of fields (attributes,variables);so,the dimensionality of the problem is high. A high-dimensional data set creates problems in terms of increasing the size of the search space for model induction in a combinatorially explosive manner. In addition, it increases the chances that a data-mining algorithmwill find spurious patterns that are not valid in general. Approaches to this problem includemethods to reduce the effective dimensionality of the problem and the use of prior knowledge to identify irrelevant variables.

Overfitting: When the algorithm searches for the best parameters for one particular model using a limited set of data, it can model not only the general patterns in the data but also any noisespecific to the data set, resulting in poor performance of the model on test data. Possible solutions include cross-vali-dation, regularization, and other sophisticated statisticalstrategies.

Assessing of statistical significance: A problem (related to overfitting) occurs when the system is searching over many possible models. For example, if a system tests model sat the 0.001 significance level, then on average, with purely random data,N/1000 of these models will be accepted as significant.

This point is frequently missed by many initial attempts at KDD. One way to deal with this problem is to use methods that adjust the test statistic as a function of the search, for example, Bonferroniad just ments for independent tests or randomization testing.

Changing data and knowledge: Rapidly changing (nonstationary) data can make pre- viously discovered patterns invalid. In addition, the variables measured in a given application database can be modified, deleted, or augmented with new measurements over time. Possible solutions include incremental methods for updating the patterns and treating change as an opportunity for discovery by using it to cue the search for patterns of change only (Matheus, Piatetsky-Shapiro, and McNeill 1996). See also Agrawal and Psaila (1995) and Mannila, Toivonen, and Verkamo (1995).

Missing and noisy data: This problem is especially acute in business databases. U.S. census data reportedly have error rates as great as 20 percent in some fields. Important attributes can be missing if the database was not designed with discovery in mind. Possible solutions include more sophisticated statistical strategies to identify hidden variables and dependencies (Heckerman 1996; Smyth et al. 1996).

Complex relationships between fields: Hierarchically structured attributes or values, relations between attributes, and more so- phisticated means for representing knowledge about the contents of a database will require algorithms that can effectively use such information. Historically, data-mining algorithms have been developed for simple attribute-value records, although new techniques for deriving relations between variables are being developed (Dzeroski 1996; Djoko, Cook, and Holder 1995).

Understandability of patterns: In many applications, it is important to make the dis- coveries more understandable by humans. Possible solutions include graphic representations (Buntine 1996; Heckerman 1996), rule structuring, natural language generation, and techniques for visualization of data and knowledge. Rule-refinement strategies (for ex- ample, Major and Mangano [1995]) can beused to address a related problem: The discovered knowledge might be implicitly or explicitly redundant.

References

- Agrawal, R., and Psaila, G. 1995. Active DataMining. In Proceedings of the First International Conference on Knowledge Discovery and Data Mining(KDD-95), 3–8. Menlo Park, Calif.: American Association for Artificial Intelligence.
- Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; and Verkamo, I.1996. Fast Discovery of Association Rules. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P.Smyth, and R. Uthurusamy, 307–328. Menlo Park,Calif.: AAAI Press.
- Apte, C., and Hong, S. J. 1996. PredictingEquityReturns from Securities Data with Minimal Rule Generation. In Advances in Knowledge Discovery and Data Mining, eds.
- U. Fayyad, G. Piatetsky-Shapiro, P.Smyth, and R. Uthurusamy, 514–560. Menlo Park,Calif.: AAAI Press.
- Basseville, M., and Nikiforov, I. V. 1993. Detectionof Abrupt Changes: Theory and Application. Englewood Cliffs, N.J.: Prentice Hall.
- Berndt,D.,andClifford,J.1996.FindingPatternsin Time Series: A Dynamic Programming Approach. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G.Piatetsky- Shapiro, P.Smyth, and R. Uthurusamy, 229–248. Menlo Park, Calif.: AAAIPress.
- Berry, J. 1994. Database Marketing. Business Week, September 5, 56–62.
- Brachman, R., and Anand, T. 1996. The Process of Knowledge Discovery in Databases: A Human-Cen-tered Approach. In Advances in Knowledge Discovery and Data Mining, 37–58, eds. U. Fayyad, G. Piatet-sky-Shapiro, P. Smyth, and R. Uthurusamy. MenloPark, Calif.: AAAI Press.
- Breiman, L.; Friedman, J. H.; Olshen, R. A.; andStone, C.J.1984. Classification and Regression Trees. Belmont, Calif.: Wadsworth.
- Brodley, C. E., and Smyth, P. 1996. Applying Classification Algorithmsin Practice. Statistics and Computing. Forth coming.
- Buntine, W. 1996. Graphical Models for Discovering Knowledge. In Advances in Knowledge Discovery and Data Mining, eds.
- U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 59–82.Menlo Park, Calif.: AAAI Press.
- Cheeseman, P. 1990. On Finding the MostProbableModel. In Computational Models of Scientific Discovery and Theory Formation, eds. J. Shrager and P. Langley, 73–95. San Francisco, Calif.: Morgan Kaufmann.
- Cheeseman, P., and Stutz, J. 1996. Bayesian Classification (AUTOCLASS): Theory and Results. In Advances in Knowledge Discoveryand Data Mining,eds.
- U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R.Uthurusamy, 73–95. Menlo Park, Calif.: AAAI Press.
- Cheng, B., and Titterington, D. M. 1994. NeuralNetworks—A Review from a Statistical Perspective. *Statistical Science* 9(1): 2–30.
- Codd, E. F. 1993. Providing OLAP (On-Line AnalyticalProcessing)toUser-Analysts: An IT Mandate.E. F. Codd and Associates.
- Dasarathy, B. V. 1991. Nearest Neighbor (NN)Norms: NN Pattern Classification Techniques. Washington, D.C.: IEEE Computer Society.
- Djoko, S.; Cook, D.; and Holder, L. 1995. Analyzingthe Benefits of Domain Knowledge in Substructure Discovery. In Proceedings of KDD-95: First International Conference on Knowledge Discovery and Data Mining, 75–80. Menlo Park, Calif.: American Association for Artificial Intelligence.
- Dzeroski, S.1996. Inductive Logic Programming for Knowledge Discovery in Databases. In *Advances in Knowledge Discovery and Data Mining*, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 59–82. Menlo Park, Calif.: AAAI Press.
- Elder, J., and Pregibon, D. 1996. A Statistical Perspective on KDD. In *Advances inKnowledge Discovery and Data Mining*, eds.
- U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 83–116.Menlo Park,Calif.: AAAI Press.
- Etzioni, O.1996. The World Wide Web:Quagmire or Gold Mine? *Communications of the ACM* (Special Issue on Data Mining). November 1996. Forthcoming.
- Fayyad, U. M.; Djorgovski, S. G.; and Weir, N. 1996.From Digitized Images to On-Line Catalogs: Data Mining a Sky Survey. *AI Magazine* 17(2): 51–66.
- Fayyad, U. M.; Haussler, D.; and Stolorz, Z. 1996.KDD for Science Data Analysis: Issues and Examples. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 50–56. Menlo Park, Calif.: American Association for Artificial Intelligence.

- Fayyad, U. M.; Piatetsky-Shapiro, G.; and Smyth, P.1996. From Data Mining to Knowledge Discovery:An Overview. In *Advances in Knowledge Discovery and Data Mining,* eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1–30. Menlo Park, Calif.: AAAI Press.
- Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; andUthurusamy, R. 1996. *Advances in Knowledge Discovery and Data Mining.* Menlo Park, Calif.: AAAIPress.
- Friedman, J.H.1989. Multivariate Adaptive Regression Splines. *Annals of Statistics* 19:1– 141.
- Geman, S.; Bienenstock, E.; and Doursat, R. 1992.Neural Networks and the Bias/Variance Dilemma. *Neural Computation* 4:1–58.
- Glymour, C.; Madigan, D.; Pregibon, D.; andSmyth, P. 1996. Statistics and Data Mining.*Communications of the ACM* (Special Issueon Data Mining). November 1996.Forthcoming. Glymour,C.;Scheines,R.;Spirtes,P.;Kelly,K.1987.
- *Discovering Causal Structure.* New York: Academic. Guyon, O.; Matic,N.;and Vapnik, N.1996. Discovering Informative Patterns and Data Cleaning. In *Advances in Knowledge Discovery and Data Mining,* eds. U. Fayyad,
- G. Piatetsky-Shapiro, P.Smyth, and R. Uthurusamy, 181–204. Menlo Park, Calif.: AAAIPress.
- Hall, J.; Mani, G.; and Barr, D. 1996. Applying Computational Intelligence to the Investment Process. In Proceedings of CIFER-96: Computational Intelligence in Financial Engineering. Washington, D.C.: IEEE Computer Society.
- Hand, D. J. 1994. Deconstructing Statistical Questions. *Journal of the Royal Statistical Society A.* 157(3):317–356. Hand, D.J. 1981. *Discrimination and Classification.* Chichester, U.K.:Wiley.
- Heckerman, D.1996. Bayesian Networks for Knowledge Discovery. In *Advances in Knowledge Discovery and Data Mining,* eds.
- U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 273–306.Menlo Park, Calif.: AAAI Press.
- Hernandez, M., and Stolfo, S. 1995. The MERGE-PURGE Problem for Large Databases. In Proceedings of the 1995 ACM- SIGMOD Conference, 127–138.New York:Association for Computing Machinery.
- Holsheimer, M.; Kersten, M. L.; Mannila, H.; and Toivonen, H. 1996. Data Surveyor: Searching the Nuggets in Parallel. In *Advances in Knowledge Discovery and Data Mining,* eds. U. Fayyad, G. Piatet-sky- Shapiro, P. Smyth, and R. Uthurusamy,447–471. Menlo Park, Calif.: AAAI Press.
- Horvitz, E., and Jensen, F. 1996. *Proceedings of theTwelfth Conference of Uncertainty in Artificial Intelligence.* San Mateo, Calif.:Morgan Kaufmann.
- Jain, A. K., and Dubes, R. C. 1988. *Algorithms for Clustering Data.* Englewood Cliffs, N.J.: Prentice-Hall.
- Kloesgen, W. 1996. A Multipattern andMultistrategy Discovery Assistant. In *Advances in Knowledge Discovery and Data Mining,* eds. U. Fayyad, G. Piatet-sky- Shapiro, P. Smyth, and R. Uthurusamy, 249–271. Menlo Park, Calif.: AAAI Press.
- Kloesgen, W., and Zytkow, J. 1996. Knowledge Discovery in Databases Terminology. In *Advances in Knowledge Discovery and Data Mining,* eds. U. Fayyad,
- G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 569–588. Menlo Park, Calif.: AAAI Press.
- Kolodner, J. 1993. *Case-Based Reasoning.* San Francisco, Calif.: Morgan Kaufmann.
- Langley, P., and Simon, H. A. 1995. Applications of Machine Learning and Rule Induction. *Communications of the ACM* 38:55–64.